Escreva sua pergunta aqui: https://app.sli.do/event/myhjiwoc

MAC0459/MAC5865 - Tópicos em Ciência e Engenharia de Dados

Aula 01

Sejam bem-vindas, sejam bem-vindos!

Entre no link https://app.sli.do/event/myhjiwoc ou e faça suas perguntas da aula.

R. Hirata Jr.

Escreva sua pergunta aqui: https://app.sli.do/event/myhjiwoc

Objetivos de hoje

- Ao final da aula de hoje você deve:
  - Saber a diferença entre:
    - o método do cientista
    - o método do engenheiro
    - o método do cientista de dados
  - Conhecer os papéis, ou trabalhos, em Ciência de Dados
  - Ter consciência de algumas das dificuldades para sua transformação para um DS

Relembrando a aula passada

Escreva sua pergunta aqui: https://app.sli.do/event/myhjiwoc

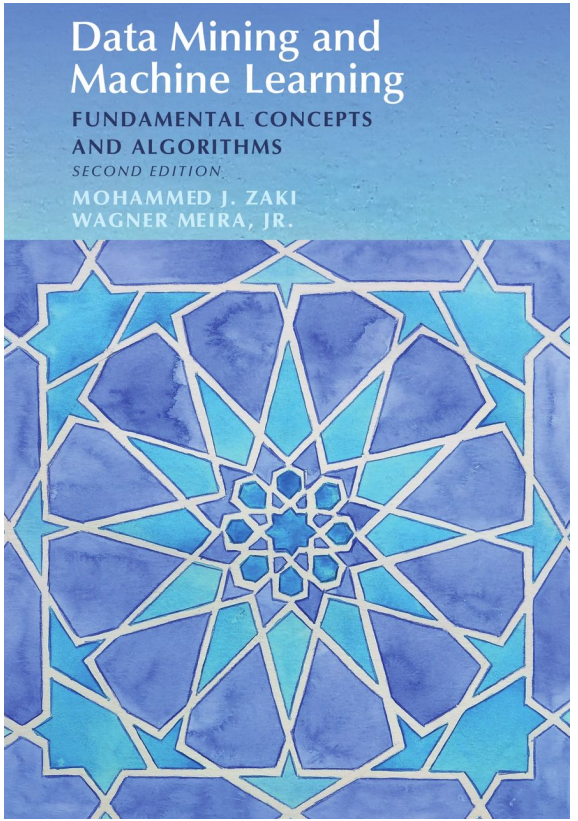
Referência principal

- The Data Science Design Manual
  - http://www.data-manual.com/
  - What is Data Science?

no e-disciplinas

Outras referências

- Data Mining and Machine Learning
- https://dataminingbook.info/
- Palestra do prof. Meira na disciplina de MAC5865
  - http://iptv.usp.br/portal/video.action?idItem=22508



Escreva sua pergunta aqui: https://app.sli.do/event/myhjiwoc

Ciência, Engenharia e DS

Simple pipeline – Scientific Method

1. Pose a question
2. Formulate a hypothesis
3. Formulate an experiment
4. Observe (data collecting)
5. Analyse the results
6. Go back to step 2 if the hypothesis is not correct/supported
7. Report results

Simple pipeline – Engineering Method

1. Define a problem
2. Specify requirements
3. Brainstorm, evaluate, choose solution
4. Develop a prototype solution
5. Tests solution
6. Go back to step 3 if the results, or data, etc in case the solution does not meet requirements

Simple pipeline – Data Science Method

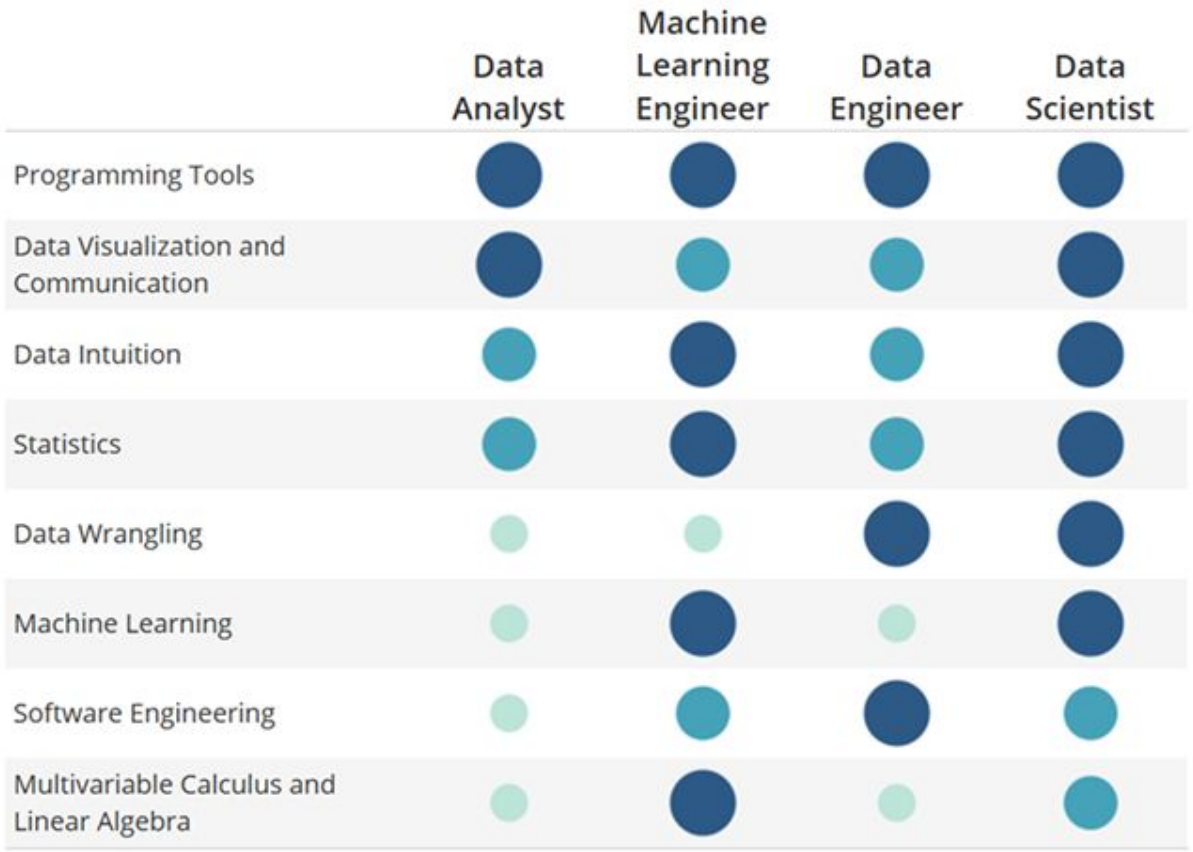
1. Pose question
2. Get the data
3. Explore the data
4. Model the data
5. Report results

Escreva sua pergunta aqui: https://app.sli.do/event/myhjiwoc

Data Science Jobs

Escreva sua pergunta aqui: https://app.sli.do/event/myhjiwoc

Data Science Jobs



https://blog.udacity.com/2018/01/4-types-data-science-jobs.html

Escreva sua pergunta aqui: https://app.sli.do/event/myhjiwoc

Pronto, pode acordar!

INSTITUTO DE MATEMÁTICA E ESTATÍSTICA
IME 50
UNIVERSIDADE DE SÃO PAULO

Escreva sua pergunta aqui: https://app.sli.do/event/myhjiwoc

Você está presente?

What is Data Science (DS)?

- The purpose of computing is insight, not numbers.
  - Richard H. Hamming
- Stop teaching calculating, start teaching math!
  - Stephen Wolfram
- The Data Science Design Manual by Steven Skiena

DS - reasons of success

- New technologies to sensor the world
- Computing advances
- Prominent tech companies
- Open source
- Public datasets
- Reproducibility/repeatability
- International conferences

Repeatability vs Reproducibility

- Repeatability
  - variability caused by the measurement device
- Reproducibility
  - variability caused by different labs/operators

Escreva sua pergunta aqui: https://app.sli.do/event/myhjiwoc

Big Data

International Conference on Knowledge Discovery and Data Mining
“Conferência Internacional em Descoberta de Conhecimento e Mineração de Dados”

CS, DS, Real Science

- “Computer scientists, by nature, don’t respect data”
- “data is just meat to be passed through a sausage grinder”
- What the number means vs what the number is (type)
- Examples?

CS, DS, Real Science

- Think like a real scientist
- Scientists obsess about discovering thing, while computer scientists invent rather than discover
- Data vs method centrism
- Concerns about results
- Robustness
- Precision

Data vs method centrism

- Scientists are data driven
- Computer scientists are algorithm driven
- Scientists invent measuring devices
- Computer scientists invent and enhance algorithms

Concern about results

- Real scientists worry about answers.
- Good scientists care about whether the results make sense
- Bad computer scientists worry about producing plausible-looking numbers

Robustness and Precision

- Real scientists are comfortable that data has errors, computer scientists in general are not.
- Nothing is ever completely true or false in science, while everything is either true or false in CS or Math

Precision

- Computer scientists are happy printing floating point numbers to as many digits as possible.
- Real scientists will only use as many significant digits as the worst precision of any measurement in the process.

Hypothesis vs data driven science

- HD science - ask specific questions of the world and then generating the specific data to confirm, or deny it
- Data driven science - a new paradigm to model the world

Hypothesis vs data driven science

- HD science:
  - given a problem, what available data will help us answer it?
- Data driven science:
  - given data, what interesting problems can we apply it to?

Escreva sua pergunta aqui: https://app.sli.do/event/myhjiwoc

Scientific questions

- To call in the statistician after the experiment is done may be no more than asking him/her to perform a postmortem examination: she/he may be able to say what the experiment died of.

Sir Ronald Fisher

Escreva sua pergunta aqui: https://app.sli.do/event/myhjiwoc

Scientific questions

Good experiments are designed

Hypothesis vs data driven science

- HD science:
  - given a problem, what available data will help us answer it?
- Data driven science:
  - given data, what interesting problems can we apply it to?

Escreva sua pergunta aqui: https://app.sli.do/event/myhjiwoc

Learning to ask questions

- Computer scientists students are not used to ask questions, why?
- Good data scientists develop an inherent curiosity about the world around them and have wide-ranging interests.

Simple pipeline – Scientific Method

1. Pose a question
2. Formulate a hypothesis
3. Formulate an experiment
4. Observe (data collecting)
5. Analyse the results
6. Go back to step 2 if the hypothesis is not correct/supported
7. Report results

Example of application

1. Pose a question
   - Is lettuce mostly composed by water?
2. Formulate a hypothesis
   - Lettuce leaves have about 95% of water
3. Formulate an experiment

Escreva sua pergunta aqui: https://app.sli.do/event/myhjiwoc

Experimental protocol

1. Food dehydrator
2. Lettuce (origin etc)
3. Clean and dry leaves
4. Weight the leaves
5. Put it to dry for 60 minutes

https://www.westonsupply.com/Weston-Stainless-Steel-Food-Dehydrator-p/74-1001-w.htm

Escreva sua pergunta aqui: https://app.sli.do/event/myhjiwoc

Experimental protocol

4. Weight the dried leaves
5. Take the difference of weights and check if it is 95% of the original weight
6. Go back to step 2 if the hypothesis is not correct/supported
7. Report results

https://www.westonsupply.com/Weston-Stainless-Steel-Food-Dehydrator-p/74-1001-w.htm

Escreva sua pergunta aqui: https://app.sli.do/event/myhjiwoc

What is Science?

- “We absolutely must leave room for doubt or there is no progress and there is no learning. There is no learning without having to pose a question. And a question requires doubt. People search for certainty. But there is no certainty. People are terrified — how can you live and not know? It is not odd at all. You only think you know, as a matter of fact. And most of your actions are based on incomplete knowledge and you really don’t know what it is all about, or what the purpose of the world is, or know a great deal of other things. It is possible to live and not know.” Feynman

Escreva sua pergunta aqui: https://app.sli.do/event/myhjiwoc

What is Science?

- In our example, did we left room for doubt?
- The lemma of our disciple should be:

De omnibus dubitandum

Learning to ask questions

- The baseball encyclopedia
- The Internet Movie Database (IMDb)
- Google Ngrams
- New York Taxi Records

The baseball encyclopedia

- How can we best measure an individual player's skill or value?
- How fairly do trades between teams generally work out?
- What is the general trajectory of player's performance level as they mature and age?

Escreva sua pergunta aqui: https://app.sli.do/event/myhjiwoc

The baseball encyclopedia

- Do left-handed people have shorter lifespans than right-handers?
- How often do people return to live in the same place where were born?
- Do player salaries generally reflect past, present or future performance?
- To what extent have heights and weights been increasing in

INSTITUTO DE MATEMÁTICA E ESTATÍSTICA
IME 50
UNIVERSIDADE DE SÃO PAULO

Escreva sua pergunta aqui: https://app.sli.do/event/myhjiwoc

Obrigado!